Library Resources & Technical Services, Vol.37, No.3, 1993, pp.59-71
ISSN: 0024-2527
http://www.ala.org
http://www.ala.org/ala/mgrps/divs/alcts/index.cfm
http://www.ala.org/ala/mgrps/divs/alcts/resources/lrts/index.cfm

Characteristics of Duplicate Records in OCLC's Online Union Catalog

Edward T. O'Neill, Sally A. Rogers, and W. Michael Oskins
Edward T. O'Neill is Consulting Research Scientist, and W. Michael Oskins is Consulting System Analyst, OCLC Online Computer Library Center, Inc., Dublin, Ohio; Sally A. Rogers is Head of Serial, Nonbook, and Thesis Cataloging, The Ohio State University Libraries, Columbus.

Duplicate records in the Online Union Catalog of the OCLC Online Computer Library Center, Inc., were analyzed. Bibliographic elements comprise information found in one or more fields of a bibliographic record; e.g., the author element comprises the main and added author entry fields. Bibliographic element mismatches in duplicate record pairs were considered relative to the number of records in which each element was present. When a single element differed in a duplicate record pair, that element was most often publication date. This finding shows that a difference in the date of publication is not a reliable indicator of bibliographic uniqueness. General cataloging and data entry patterns such as variations in title transcription and form of name, typographical errors, mistagged fields, misplaced subfield codes, omissions, and inconsistencies between fixed and variable fields often caused records that were duplicates to appear different. These factors can make it extremely difficult for catalogers to retrieve existing bibliographic records and thus avoid creating duplicate records. They also prevent duplicate detection algorithms used for tape-loading records from achieving desired results. An awareness of particularly problematic bibliographic elements and general factors contributing to the creation of duplicate records should help catalogers identify and accept existing records more often. This awareness should also help to direct system designers in their development of more sensitive algorithms to be used for tape loading. The resulting general reduction in the number of duplicate records in union catalogs will be a major step toward increased cataloger productivity, user satisfaction, and overall online database quality.

Definition and Statement of Problem

Duplicate records that are identical are rare and easy to identify in union catalogs. As Hunstad (1988, 246) notes when discussing the problem of duplicates:

> A procedure that matches two bibliographic descriptions byte by byte, and says yes, when all bytes match in number and in kind, and no when any slight difference occurs, will be simple enough to make, but will not fight actual duplication in a database.

It is the duplicate records that are similar but not identical that are hard to identify. Duplicate record studies and detection processes are useful only if they focus on these similar records. Duplicate records in the Online Union Catalog of the OCLC Online Computer Library Center, Inc., originate from batch loading of bibliographic records and from online cataloging. In a shared cataloging database, changes in cataloging rules and variations in cataloging practices can result in significantly different records for the same bibliographic item. In the case of batch loading, duplicate records are created when variations between the records being loaded and the corresponding records in the database prevent the system from identifying the duplicates. Duplicate records are created through online cataloging when a cataloger does not find the existing database record or does not recognize it as representing the item to be cataloged.

Failure to find an existing record can be due to inability or unwillingness to thoroughly search the database. Derived keys may be inadequate to retrieve an existing record. Improper searching techniques or errors in a bibliographic record can also prevent retrieval. Failure to recognize a record as representing the item to be cataloged can occur when the record includes more, less, or different information than appears on the item. This situation often results from a lack of uniformity among publishers in the use of publication date, printing date, and edition statements.

Duplicate records in online shared catalogs reduce efficiency in proportion to the actual number of records in the database. The greater the number of duplicate records online, the more difficult it is to search the database and to make cataloging decisions. Duplicate records also have an impact on library automation cost-effectiveness through indirect expenditures of time (e.g., searches and redundant cataloging) and direct expenditures of fees (e.g., searches; interlibrary loan; and EPIC, OCLC's online reference service).

Duplicate Record Sample

A random sample from the Online Union Catalog was used to characterize duplicate records. The sample was formed by randomly generating OCLC control numbers. When the randomly generated numbers matched a bibliographic record for a book (bibliographic level = m, type of record = a), the record was included in the sample. Professional catalogers then manually searched the Online Union Catalog to identify as many potential duplicates as possible for each record in the sample. Automated searching techniques, developed to identify duplicate records in the Online Union Catalog were also used to identify additional potential duplicate records (Campbell 1991, 24). The results of both the manual and automated searches were combined into a single set of possible duplicates.

Thirteen specific bibliographic elements were used as a means of identifying and analyzing variations to determine record duplication. The elements used for this study included title, author, media (i.e., microreproduction or photocopy), date, control numbers (International Standard Book Number [ISBN], Library of Congress Card Number [LCCN], and Government Documents Control Number [GDCN]), statement of responsibility, edition, publisher, pages, size, and series. The use of elements rather than specific fields is an important distinction: Elements are descriptive of a like kind of information about a book and may include data found in more than one fixed or variable field of a bibliographic record. For example, the media element combines data from the fixed field Repr (form of reproduction) as well as the variable fields 007 (physical description), 500 (general note), and 533 (reproduction note).

Each pair of potential duplicates was evaluated by three to seven professional catalogers. Records were classified as either duplicates, nonduplicates, or unknown, pending physical examination of the books they represented. More than one hundred of these unknown duplicates were obtained through interlibrary loan and reviewed physically. The result of all of these evaluations yielded the 742 confirmed duplicate record pairs analyzed in this study. All of the examples in this paper were taken from the 742 duplicate records pairs.

Because the sample was derived solely from the Online Union Catalog, no statistical basis exists for extrapolating to union catalogs in general. However, the types of problems observed in this study are likely to be found in other union catalogs.

General Characteristics of Duplicate Records

A review of the confirmed duplicate records identified several general cataloging and

data entry factors that contribute to duplication in the Online Union Catalog. These factors include (1) typographical errors, (2) erroneous tags and subfield codes, (3) omitted information, and (4) inconsistencies between the variable and fixed fields. Examples of these and other characteristics of duplicate records are included below and are reproduced exactly as they appeared on catalog records in the Online Union Catalog. Consequently, the examples might not reflect correct cataloging or the application of current rules and standards.

Typographical Errors

Typographical errors were found throughout the sample of duplicates. Although the error rate was relatively low, typographical errors in critical elements such as author, title, or control number can make it difficult to retrieve records.

As addressed within this study, transcription errors included spelling and transcription errors, as well as keying errors. Differences in capitalization, spacing, punctuation, and diacritic usage were not considered significant and were not counted as typographical errors.

Most typographical errors were one-character differences, e.g., the substitution of c for s in the title:

The great school bus controversy
The great school bus controvercy

or the addition of an *i* in the publisher:

Macmillan
Macmillian

In another example, unlawful was dropped from the title:

Investigation of arson, and other unlawful burnings
Investigation of arson and other burnings

While typographical errors were found in all fields, they were most common in the longer textual fields, particularly author, title, and publisher.

Erroneous Tags and Subfield Codes

Mistagged fields and erroneous or misplaced subfield codes were another common problem. Some examples included (a) an author transformed into a subject by tagging the field 600 rather than 700, (b) statements of responsibility coded as subtitles, and (c) publishers coded as the place of publication. The title entry

The Federalist or the new constitution by Alexander Hamilton, James Madison & John Jay

is a good example. That particular title was found with a ≠ b subfield code following The Federalist but without a ≠ c subfield code, creating the subtitle

or the new constitution by Alexander Hamilton

Erroneous or misplaced subfield codes often result in fields that algorithmically contain different information even when the contents of the fields are otherwise identical.

Omitted Information

Initially, catalogers reviewing potential duplicate records tended to be conservative and to assume that information not present on a bibliographic record was not present on the book it represented. For example, if one record included a series statement and a potential duplicate record did not, the omission was initially thought to be an indication that different editions of the title had been published. However, as more records were reviewed — particularly those that

were paired with multiple potential duplicates — after examining the actual books, it became increasingly clear that omitted information usually was not significant. Omitted information was not limited to less-than-full-level records. Less-than-full-level records were not significantly more likely to be duplicates than were full-level records.

Analysis of duplicate record pairs in the sample showed that the omission of title information, author entries, edition statements, secondary publishers or distributors, and series statements did not necessarily mean that this information was not on the piece. In fact, in many cases the element was present, but some catalogers were more selective than others in deciding what to include in the bibliographic description, even when information appeared on the title page. These inconsistencies undoubtedly reflect differing local needs and a practice of including in the record only those elements of the description considered important for local use.

Inconsistencies between the Variable and Fixed Fields

Fixed fields were included in the Machine-Readable Cataloging (MARC) record to ease automated processing, and in fact, most duplicate detection algorithms rely on the use of fixed fields for this reason. However, reliance on algorithms of fixed fields and their exclusion of the use of variable fields limit their accuracy in identifying duplicate records. This primary use of fixed fields also assumes greater reliability in the fixed fields. This is not necessarily the case, even though fixed fields contain simple encoded information. In fact, variable fields can provide more reliability because they contain the originating information; i.e., fixed fields contain a coded version of free-text information from the variable fields. Because the fixed fields do not contain original source information, they present a chance for error in encoding information and little, if any, chance for automatic correction. A one-character typographic error in a fixed field results in an immediate mismatch between potential duplicate records. For example, in one of the duplicate record pairs, the country of publication code did not match because one record had Connecticut coded correctly as *ctu*, but the other record had it coded as *cnu*.

Because variable fields are free-text, they provide for some natural redundancy within themselves. Differences can be considered within the context of surrounding letters or words and do not necessarily result in mismatches, especially with manual review. For example, in one duplicate pair, country of publication information differed between the variable fields (260 ≠a subfield) but not between the fixed fields. The first record identified the place of publication as *Allahabad* while the second record had *Allahabad (India)*. While *Allahabad* and *Allahabad (India)* would certainly be recognized as equivalent during a manual review, developing software with the ability to recognize these as the same information is difficult. Comparing the three-character code in the fixed field is, by contrast, straightforward.

The most common inconsistency in the form of reproduction code in the fixed field of potential duplicate records was leaving the code blank even when there was a clear indication in the variable fields that the monograph was a microform. Because a blank value in this field has meaning and, in fact, is the default value, it may represent a purposeful selection (meaning the item is not a reproduction), a miscoding, or an omission. Therefore, data from the variable fields (e.g., *533*) were included in the media element to improve the accuracy of distinguishing records for reproductions from records for the originals.

While fixed fields are easier to use for computer matching and standardization of information, the likelihood of critical, unforgiving errors requires that the fixed fields be used in conjunction with variable fields if duplicates are to be reliably identified. When variable fields express the same information in different ways (e.g., place of publication), fixed fields can be

used to obtain the record match. If the fixed fields do not match, variable fields can be checked so that errors in fixed fields become less crucial.

Notes Fields

During manual review, information in the general notes fields of potential duplicate records often provided the clearest indication of whether records were duplicates. In one potential duplicate record, changes in the content of an item distinguishing it from an earlier edition were described in a note, although the record lacked a formal edition statement. In other examples, indications that a record represented a reproduction and not the original were included in a general note rather than in designated areas; e.g., fixed fields, reproduction note. A quoted note in one record often contained information found in a subtitle or series statement in another record.

According to the *Anglo-American Cataloguing Rules*, 2d ed., 1988 revision (AACR2R), "Notes contain useful descriptive information that cannot be fitted into other areas of the description" (Gorman and Winkler 1988, 50). Because notes are entered in free form, however, their usefulness in duplication detection processes is generally limited.

Bibliographic Characteristics of Duplicate Records

In addition to general factors such as typographical errors, erroneous tags/subfield codes, omissions, and fixed/variable field inconsistencies, each of the thirteen bibliographic elements used to evaluate potential duplicate records had its own unique problems. These problems could be attributed to changing cataloging rules, differing interpretations of the rules, local practice, or simply misinterpretation.

As indicated in figure 1, duplicate pairs frequently had more than one bibliographic element that differed between the records. A total of 28% of the duplicates had two elements that were different, and another 22% had mismatches in three elements. Twenty-six percent of the duplicate pairs differed in a single bibliographic element. These differences were most frequently due to differing cataloging interpretations of descriptive information for an edition.

When looking at duplicates that had only one element that was different, the date, author, and publisher elements were the ones that most frequently did not match between records. When individual elements were analyzed in duplicates with more than one mismatch, the elements that differed between duplicate records, in order of frequency, were date (58%), author (33%), publisher (29%), title (23%), size (23%), pages (23%), statement of responsibility (19%), and series (18%). Single-element and multiple-element mismatches as characteristics of duplicate records are compared in figure 2.

Key elements of the description that differentiated potential duplicate records included pagination, particularly preliminary paging (roman numerals), and statement of responsibility, especially when phrases such as with new introduction explained differences from other editions.

The commonality of bibliographic elements in duplicates is highlighted in figure 3. Title, media, and date elements were most frequently present in duplicates. Least common were the control numbers, edition, and series elements.

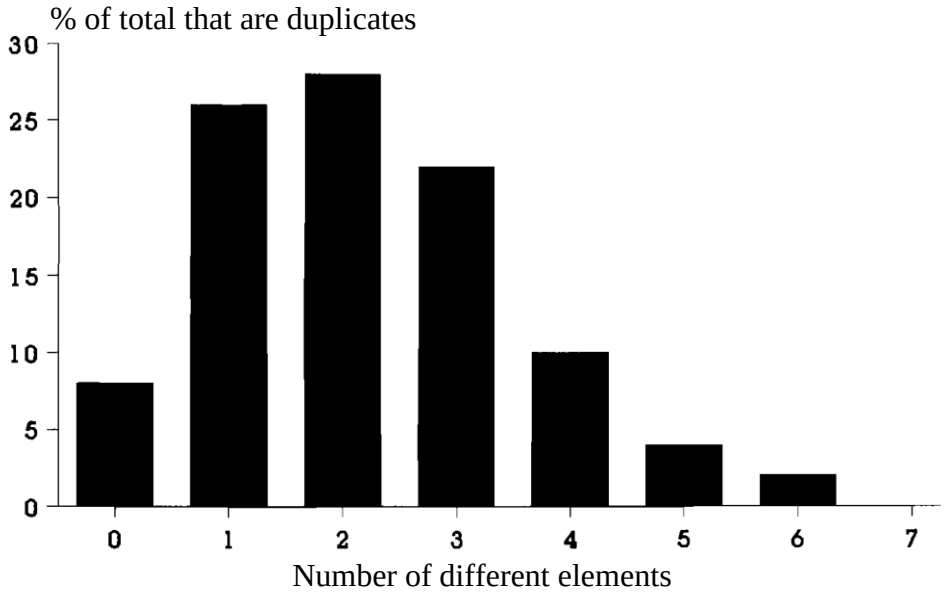

Figure 1. Duplicate Records: Number of Elements Present.

Author Element

More than 30% of the duplicates had differing author entries. These differences tended to be insignificant, occurring primarily in the form of the entry or, in the case of multiple authors, the determination of which authors to include. Settlement Patterns of the Western Hueco Bolson, whose title page is shown in figure 4, exemplifies one of the duplicate pairs with differing author entries.

For the first record in the duplicate pair, the author entries were

100 10 *Whalen, Michael E.*

710 10 *United States.* ≠b *Army.* ≠b *Corps of Engineers.* ≠b *Fort Worth District*

In the second record, the entries were

100 10 *Whalen, Michael Edward,* ≠d 1948-

700 10 *Gerald, Rex E.*

The Library of Congress authority file entry for this author is *Whalen, Michael E.* This example is typical of the variations found in the author element in duplicate records.

Differing forms of entry are characteristic of duplicate records when the differences are in the author element. In the preceding example, one cataloger located and used the author's full name and date of birth; the other cataloger used the name as it appeared on the title page. Depending on when each of the records was created, this example might also illustrate a change in the cataloging rules concerning fullness of name in author entries. Differences in the forms of entry may also result from changes in the Library of Congress authority file.

Catalogers also have a choice as to which added entries they make. In the preceding example, one cataloger preferred the *Corps of Engineers*, for whom the report was prepared; the other selected *Gerald, Rex E.*, the principal investigator.

Other author characteristics in duplicate records included authors publishing under different names (i.e., name changes, pseudonyms), differing forms of entry for corporate names, and the selection of a personal or a corporate name for the main entry rather than an added entry

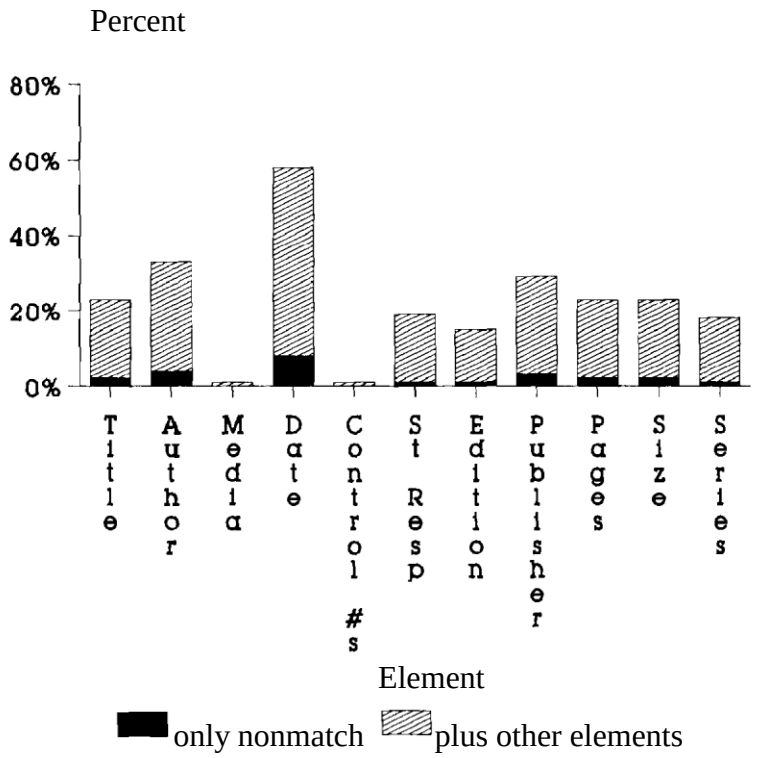

Figure 2. Duplicate Records: Non-Matching Elements.

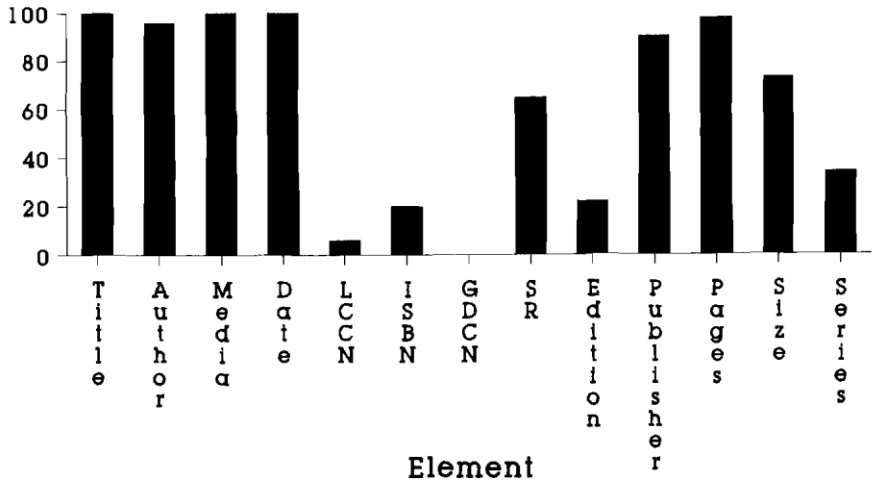

Figure 3. Duplicate Records: Percent of Time Field is Present.

Title Page

SETTLEMENT PATTERNS OF THE WESTERN HUECO BOLSON
BY
MICHAEL E. WHALEN, PH.D.
PROJECT ARCHEOLOGIST

WITH CONTRIBUTIONS BY

T. C. O'LAUGHLIN. M.S., J. D. PIGOTT. M.S., C. C. STOUT, PH.D.
M. K. STOUT, A.I.A., AND W. E. WETTERSTROM, PH.D.

REX E. GERALD, PH.D.
PRINCIPCAL INVESTIGATOR

PREPARED FOR
THE CORPS OF ENGINEERS
FORT WORTH DISTRICT, FORT WORTH, TEXAS

UNDER CONTRACT DACA63-76-C-0219

HISTORIC AND NATURAL RESOURCES REPORT
U.S. ARMY, FORT BLISS
ENVIRONMENTAL OFFICE, DFAE
EL PASO, TEXAS

PUBLICATIONS IN ANTHROPOLOGY NO. 6
EL PASO CENTENNIAL MUSEUM
THE UNIVERSITY OF TEXAS AT EL PASO

JULY 1978

Figure 4. Replica of Title Page from a Duplicate Record Pair with Differing Author Entries.

Title and Statement of Responsibility Elements

Consistent identification of a title and/or statement of responsibility can be difficult. Aside from the typographical and subfield coding errors discussed previously in this paper, one of the most common problems found in duplicate records is different interpretations of what constitutes the title and how to transcribe it. In the following four examples, the second title is fuller than the first; thus, the titles did not match in duplicate record pairs.

1. *Swedish Catalogue*
   Swedish Catalogue; International Exhibition, 1876
2. *Guide to Company Balance Sheets and Profit Loss Accounts*
   Guide to Company Balance Sheets and Profit Loss Accounts, for Directors, Secretaries, Accountants, Bankers, Investors, and Students
3. *The Resistance to Impact of Spent Magnox Fuel Transport Flasks*
   The Resistance to Impact of Spent Magnox Fuel Transport Flasks: Papers Presented

at a Seminar Held at the Institution of Mechanical Engineers on 30 April and 1 May 1985

4. *Calculus Made Easy: Being a Very Simplest Introduction to . . . the Differential Calculus and the Integral Calculus*
   Calculus Made Easy: Being a Very Simplest Introduction to Those Beautiful Methods of Reckoning Which Are Generally Called by the Terrifying Names of the Differential Calculus and the Integral Calculus

In the fourth example, text was omitted from the middle of the title and replaced with ellipses.

In other cases, even the beginnings of titles were different. For example,

Do Not Go Gentle into That Good Night

CBS Playhouse Presents, Do Not Go Gentle intoThat Good Night

Or, as in the following example, the position of the date within the title varied:

Scott 1990 Standard Postage Stamp Catalogue

Scott Standard Postage Stamp Catalogue, 1990

Differences in the title were found in more than 20% of the duplicate records. Because the title is a primary element in any catalog record, title variations pose significant problems. Of the duplicate records with differing titles, 40% had shortened titles; 30% had typographical errors in titles; and 30% had a variety of differences, including both a shortened title and typographical errors in the same duplicate pair.

Differences in the statement of responsibility also were significant, occurring in 19% of the duplicates. In addition to typographical errors, the most common inconsistencies in the statement of responsibility were due to the omission of the statement of responsibility or problems identifying the end of the statement of responsibility.

Edition and Date Elements

Jones and Kastner (1983) identified the difficulty of distinguishing printings and editions of a given monographic tide as a major factor that contributed to adding duplicate records to union catalogs. This observation is empirically confirmed by data collected in this study relative to the edition and date elements in the duplicate records.

Publication date was the only difference in 8% of the duplicates; another 50% differed in publication date and at least one other element. Of all the potential duplicates that differed only in publication date, only one represented different editions rather than printings. That is, all but one pair represented items that were "produced from essentially the same type image" and therefore constituted the same edition, according to *AACR2R* (Gorman and Winkler 1988, 617).

The edition element was present in less than 20% of the duplicate records. Of those duplicate records that included edition statements, 4% differed solely because of mismatching edition statements; 64% differed by edition and at least one other element.

Jones and Kastner provide a detailed description of the problems and considerations associated with the cataloger's need to distinguish between editions and printings. They discuss the historic eras of printing technology — hand press, machine press, and computerized printing — and emphasize the impact on cataloging. Technology has forced catalogers to distinguish between printings and editions for 19th and 20th century imprints (i.e., those produced after the hand press era).

As Jones and Kastner point out, confusion about what constitutes a different edition primarily derives from a "lack of uniformity among publishers in the use of the terms 'edition' and 'impression' or 'printing' and their equivalents in other languages" (1983,213). Cataloging

rules have acknowledged this ambiguity but still have supported the inclusion of distinguishing information in the catalog record whenever there is any doubt that the item in hand represents a distinct edition. Rules concerning publication date have added to the confusion.

One of the more prominent examples of duplicate records resulting from confusion related to editions and printings is *An Economic Interpretation of the Constitution of the United States*, title pages of which are shown in figure 5.

Earlier Printing

AN ECONOMIC INTERPRETATION OF THE CONSTITUTION OF THE UNITED STATES

BY
CHARLES A.BEARD

WITH NEW INTRODUCTION

NEW YORK
THE MACMILLAN COMPANY
1936

Later Printing

AN ECONOMIC INTERPRETATION OF THE CONSTITUTION OF THE UNITED STATES

BY
CHARLES A. BEARD

WITH NEW INTRODUCTION

NEW YORK
THE MACMILLAN COMPANY
1959

Figure 5. Replicas of Title Pages from a Duplicate Record Pair Caused by Differing Interpretations of Editions and Printings.

These are different printings of the same edition, with the later date representing a printing date (16th printing in 1959). Only one mismatched element (publication date) occurred in the records for these two items. The items represented were borrowed through interlibrary loan for physical examination, and the records were found to be duplicates. The Online Union Catalog contained 24 records representing two distinct editions of this title.

By the time Jones and Kastner wrote their article in 1983, a Library of Congress rule interpretation (LCRI) "clarified the cataloging rules in the direction of avoidance of duplication of records" (Jones and Kastner 1983, 216). This trend has continued. The LCRI for rule 1.2B4 (Library of Congress 1989) stresses that the cataloger should not supply an edition statement when one is lacking, unless differences from other editions are "manifest." The LCRI for rule

1.4G4 (Library of Congress 1989) specifically notes that the goal is for one bibliographic record to stand for all impressions.

When an edition statement on a piece to be cataloged differs from the edition statement on an existing catalog record, the cataloger must decide whether the item really represents a distinct edition or simply another printing. Rule 1.2B3 in *AACR2R* says "In case of doubt about whether a statement is an edition statement take the presence of such words as edition, issue, version, (or their equivalents in other languages) as evidence that such a statement is an edition statement" (Gorman and Winkler 1988, 30). Unfortunately, a piece that claims to be an edition often is not, particularly in the case of romance language publications.

When there is no edition statement on a piece to be cataloged and the date on the item differs from the date on an existing catalog record, the piece could be either another printing or a new edition. Although the cataloging rules describe in detail how to transcribe publication, copyright, and manufacturing dates found in an item, they do not indicate how to decide whether an undesignated date is a publication or printing date. The results of this study indicate that, in the overwhelming majority of eases, an item is just another printing when the date is the only difference.

Historic periods represented by date entries also characterize duplicate records. Duplicate records with publication dates prior to 1969 were three times more likely to have mismatches in the date field than those records with date entries after 1969. The decrease in the number of differences with imprint dates after 1969 correlates with the introduction of the use of ISBNs, which made matching of records in automated databases easier.

Control Number Elements

Fields containing the control numbers — ISBN, LCCN, and GDCN — are the most reliable variable fields. When records contain matching control numbers and have other similarities, they are usually duplicates; when they contain different numbers, they rarely are duplicates.

Errors can and do occur in control numbers; no information is immune from typographical errors. Errors in the ISBN can usually be detected because the number contains a check digit. Only one pair of duplicate records in the sample was found with differing but valid ISBNs. In that case, the book carried two ISBNs — one for hardback and one for paperback editions.

Publisher Element

Considerable variation in publisher names characterizes duplicate records. AACR2R (1.4D2) states that the publisher's name should be given "in the shortest form in which it can be understood and identified internationally" (Gorman and Winkler 1988, 37). Because there is no authority file for publisher names, it is impossible to achieve consistent results when applying this rule. Minor variations, e.g.,

Thomas B. Mosher
T.B. Mosher

are relatively common. Other publisher names are frequently shortened by abbreviation. For example:

H.M.S.O. or *H.M. Stationery Off.* for Her Majesty's Stationery Office
American Pub. Co. for the American Publishing Company
USL for U.S. League of Saving Institutions

For international databases, an additional requirement of *AACR2R* (1.4D5) was to add a subsequently named publisher if that publisher was located "in the home country of the cataloguing agency and the first named publisher ... is not" (Gorman and Winkler 1988, 38). While the logic is clear, the result is that the publisher statement is based not only on the properties of the item cataloged but also on the location of the cataloger. The publisher entries *London: J. M. Dent; New York: E. P. Dutton* and simply *London: Dent* are both correct, but they result from differing practices in American and British cataloging agencies, respectively.

In 1990, the Library of Congress issued a rule interpretation that eliminated the need to consider the location of the cataloger when recording publisher information. The LCRI calls for the inclusion of the names of all publishers on the chief source (Library of Congress 1989, 1.4D5): "Record the names of all publishers appearing on the chief source of information of the edition being cataloged.... Record also the name of a U.S. publisher appearing anywhere on the item when a non-U.S. publisher appears on the chief source."

Publisher statements also vary in the degree to which words or phrases indicating the function performed are retained. One duplicate record in the sample contained *Printed for R. Banks* in the publisher area, while the other contained only the name *R Banks*.

In some instances, it is difficult to distinguish among the author, publisher, printer, distributor, etc., resulting in another significant source of variation. The duplicate records pair for *Settlement Patterns of the Western Hueco Bolson* exemplifies this problem. In one of the records, the publisher was identified as *El Paso Centennial Museum, University of Texas*; in the other record, the publisher was identified, in brackets, as *Texas Western Press*.

The publisher is a key to determining whether two different records represent distinct editions. When a title is published by a different publisher, it is considered to be a distinct edition. A total of 29% of the duplicates in the sample were characterized by different publisher entries, posing a significant problem in identifying records as duplicates.

Series Element

The most common variations found in series entries were (1) series information was omitted from one of the records, (2) the format of the information was different, and (3) the information was recorded in different fields.

Differences frequently occurred when catalogers had to determine what was a series statement or whether a series statement was significant enough to include in the record. One book published by Charles Scribner's Sons included the Pagurian Press logo — a scorpion surrounded by a ring with the phrase "A PAGURIAN PRESS BOOK" — on the verso. One of the duplicate records for this item included the untraced series A Pagurian Press book, while the other record did not have a series statement. This omission of series information in duplicates was the most frequent difference observed for series, accounting for approximately 60% of the differences in duplicates involving series.

Even when both records contained series fields, the format of the information often differed. The following series statements taken from records in the Online Union Catalog describe the same item, although the manner in which the information is presented is very different (and not necessarily correct).

Journal of the Royal Society of Medicine.
Supplement, ≠x 0267-5331; ≠v no. 12 (1986)
Journal of the Royal Society of Medicine,
≠x 0267-5331; ≠v v. 79, suppl. no. 12

While these series statements can be recognized as equivalent with little difficulty during a manual review, the differences are difficult to overcome algorithmically.

The MARC format has nine different fields in which series information can be found. The asterisk (*) denotes pre-AACR2 forms.

400° Personal name/title series statement (traced the same)
410° Corporate name/title series statement (traced the same)
411° Conference or meeting name/title series statement (traced the same)
440 Title series statement (traced the same)
490 Series statement not traced or traced differently
800 Personal name/title series added entry
810 Corporate name/title series added entry
811 Conference or meeting name/title series added entry
830 Uniform title series added entry

This variety causes significant matching problems, particularly between pre- and post-*AACR2* forms. The duplicate records for *Innovation and Change in Reading Instruction*, the title page of which is shown in figure 6, illustrate this problem.

One of the duplicate records for this item contained a pre-AACR2 series statement in the form:

Its ≠ t Yearbook: ≠ v no. 67, ≠p part II

The equivalent series information in the other duplicate record was entered as

Yearbook of the National Society for the
Study of Education, ≠n67th, pt. 2

In addition to different entry styles, the records had different numeric styles; i.e., roman numerals and the Arabic equivalent.

Page Element

The page element is a significant characteristic of duplicate records, often being one of the more reliable indicators of whether records represent the same bibliographic item. The page element was present in 98% of the duplicates. In 2% of the duplicates, the page element was the only differing element; in 21% it was one of multiple differing elements.

Media Element

For the media element, the most common problem occurred when the Repr (form of reproduction) fixed field code was blank, despite a clear indication in a variable field that the item was a reproduction. The difficulty in determining the form of item from the variable fields is that the information may be in a variety of locations, including

007 physical description,
245 title statement, general material designation,
500 general notes as enhancements of the physical description,
533 photoreproduction note.

In most systems, items that have different codes in the Repr field are not identified as duplicates, even though the differences may be due to typographical errors and the variable field information may be indicative of a potential duplicate record. While only 1.5% of the duplicates identified in this study contained different codes, this type of error is difficult to detect.

Size Element

The physical size of the item was insignificant in characterizing duplicate records. Where duplicate records contained conflicting size data, the difference was generally only one or two centimeters. Minor differences in size occur because of inexact measurement, rounding errors, or binding differences.

INNOVATION AND CHANGE
IN READING INSTRUCTION

The Sixty-seventh Yearbook of the
National Society for the Study of Education

PART II

by
THE YEARBOOK COMMITTEE
and
ASSOCIATED CONTRIBUTORS

Edited by
HELEN M. ROBINSON

Editor for the Society
HERMAN G. RICHEY

NSSE
1968

Distributed by THE UNIVERSITY OF CHICAGO PRESS
CHICAGO, ILLINOIS 60637

Figure 6. Replica of a Title Page from a Duplicate Record Pair with Differing Series Entries.

Conclusions

This analysis of duplicate records in OCLC's Online Union Catalog demonstrates that differences between bibliographic records or between an existing record and an item to be cataloged often are not bibliographically significant. However, they often do prevent successful retrieval or matching of records, either manually or automatically. Therefore, when a bibliographic record contains minor variations from an item being cataloged, several possibilities should be considered before creating a new record. These include the possibility that the existing bibliographic record contains

1. input errors, e.g., typographical errors, or failure to change or delete fields when a new record is derived from an existing record,
2. omissions, e.g., information present on the piece may have been omitted if not needed locally,
3. different interpretations of identical information, e.g., assuming that an undesignated date is a publication rather than a printing date, or truncating a title.

Even when these possibilities are considered, it may be difficult to determine whether a record matches a piece to be cataloged. Knowing which bibliographic elements and general factors most often contribute to the creation of duplicate records should help catalogers make appropriate decisions about when to input new records. If it can be concluded that an existing record is incorrect, a change request should be submitted so that the errors may be corrected. Duplicate records should not be created on the basis of dissatisfaction with the quality of existing records. A reduction in the number of duplicate records being added to the Online Union Catalog will definitely facilitate increased cataloging productivity and end-user satisfaction, particularly with applications such as interlibrary loan and EPIC.

Works Cited

Campbell, Nancy. 1991. Database clean-up improves cataloging, ILL, and reference*OCLC newsletter* no.192 (July/Aug.).

Gorman, Michael, and Paul W. Winkler, eds. 1988. *Anglo-American cataloguing rules*, 2d ed., 1988 rev. Chicago: American Library Assn.

Hunstad, Siv. 1988. Norwegian bibliographic databases and the problem of duplicate records. *Cataloging & classification quarterly* 8, no.3/4:239-48.

Jones, Barbara, and Arno Kastner. 1983. Duplicate records in the bibliographic utilities: A historical review of the printing versus edition problem. *Library resources & technical services* 27:211-20.

Library of Congress, Office for Descriptive Cataloging Policy. 1989. *Library of Congress Rule Interpretations*. Ed. by Robert M. Hiatt. Washington, D.C.: Cataloging Distribution Service, Library of Congress.